Telling Two Distributions Apart: a Tight Characterization

Eyal Even Dar * and Mark Sandler**

Abstract. We consider the problem of distinguishing between two arbitrary black-box distributions defined over the domain $[n]$, given access to s samples from both. It is known that in the worst case $\tilde{O}(n^{2/3})$ samples is both necessary and sufficient, provided that the distributions have L_1 difference of at least ε. However, it is also known that in many cases fewer samples suffice. We identify a new parameter, that precisely controls the number of samples needed, and present an algorithm that requires the number of samples only dependent of this parameter and *independent* of the domain size. Also for a large subclass of distributions we provide a lower bound, that matches our algorithm up to the polylogarithmic bound.

1 Introduction

One of the most fundamental challenges facing modern data analysis, is to understand and infer hidden properties of the data being observed. Property testing framework [9,6,2,1,3,4] has recently emerged as one tool to test whether a given data set has certain property with high probability with only a few queries. One problem that commonly arises in applications is to test whether several sources of random samples follow the same distribution or are far apart and this is the problem we study in this paper. More specifically, suppose we are given a black box that generates independent samples of a distribution P over $[n]$, a black box the generates independent samples of a distribution Q over $[n]$, and finally a black box that generates independent samples of distribution T, which is either P or Q. How many samples do we need to decide whether T is identical to P or to Q? This problem arises regularly in change detection problems [5], testing whether Markov chain is rapidly mixing [2], and other contexts.

Our Contribution Our results generalize on the results of Batu et al in [2], they have shown that there exists a pair of distributions P, Q on domain $[n]$ with a large statistical difference $||P - Q||_1 \geq 1/2$, such that no algorithm can tell apart case (P, P) from (P, Q) with $o(n^{2/3})$ samples. They also provided an algorithm that nearly matches the lower bound for a specific pair of distributions.

In the present paper, instead of analyzing the "hardest" pair of distributions, we *characterize* the property that controls the number of samples one needs to

* evendar.eyal@gmail.com, Final Inc, this research was performed while author was at Google Research, NY

** sandler@google.com, Google Research, NY

tell a pair of distributions apart. This characterization allows us to factor out the dependency on the domain size. Namely, for every two distributions, P and Q, that satisfy certain technical properties that we describe below, we establish both an algorithm and a lower bound, such that the number of samples both necessary and sufficient is $\Omega(\frac{\|P+Q\|_2}{\|P-Q\|_2^2})$. For the lower bound example of [2], this quantity amounts to $n^{2/3}$ and the distributions in [2] satisfy the needed technical properties, and thus our results generalize upon their result.

From practical perspective such characterization is important because the high level properties of the distributions may be learned empirically (for instance it might be known that the potential class of distributions is a power-law) and our results allow to significantly reduce the number of samples needed for testing.

In many respects, our results complement those of Valiant [10]. There it was shown that for testing symmetric and continuous properties, it is both necessary and sufficient to consider only the high frequency elements. In contrast, we show that for our problem the *low* frequency elements provide all the necessary information. This was quite surprising provided that low frequency elements have no useful information for continuous properties. For the computability part [10] introduces the canonical tester - an exponential time algorithm that finds all feasible underlying data, that could have produced the output. If all checked inputs have the property value consistent, it reports it, and otherwise gives random answer. In contrast, our algorithm guarantees that for any underlying pair of distributions, the algorithm after observing the sample will be correct with high probability, even though there might be a valid input that would cause the algorithm to fail.

Finally, we develop a new technique that allows tight concentration bounds analysis of heterogeneous balls and bins problem that might be of independent interest.

Paper Overview In the next section we describe our problem in more detail, connect it to closeness problem studied in the earlier work and state our results. Section 4 proves useful concentration bounds, and introduces the main technique that we use for the algorithm analysis. Section 5 provides algorithm and analysis, and finally in the section 6 we prove our lower bounds.

2 Problem Formulation

We consider arbitrary distributions over the domain $[n]$. We assume that the only way of interacting with a distribution is through a blackbox sampling mechanism. The main problem we consider is as follows:

Problem 1 (Distinguishability problem). Given "training phase" of s samples from X and s samples from Y, and a "testing phase" of a sample of size s from either X or Y, output whether first or second distribution generated the testing sample.

We say that an algorithm solves the distinguishability problem for a class of distribution pairs $\mathcal{P}$, with s samples, if for any $(X, Y) \in \mathcal{P}$, with probability at least $1 - \frac{polylog(s)}{s^2}$ it outputs correct answer. Further, if X and Y are identical, it outputs first or second with probability at least $0.5 - \frac{polylog(s)}{s^2}$.

We show that the distringuishability problem is equivalent to the following problem studied in [2,10]:

Problem 2 (Closeness Problem[2]). Given s samples from X and s samples from Y, decide whether X and Y are almost identical or far apart with acceptable error at most $\frac{polylog(s)}{s^2}$.

An algorithm solves the closeness problem for a class of distribution pairs $\mathcal{P}$, if for every pair $(X, Y) \in \mathcal{P}$, it outputs "different" with probability at least $1 - \frac{polylog(s)}{s^2}$, and for every input of the form (X, X) it outputs "same" with probability at least $1 - \frac{polylog(s)}{s^2}$.

Our first observation is that if either of the problems can be solved for a certain class of distribution pairs, then the other can be solved with at most logarithmically more samples and time. The following lemma formalizes this statement, and due to space limitations the proof is deferred to appendix.

Lemma 1. *If there is an algorithm that solves* distinguishability *problem for a class of distribution pairs $\mathcal{P}$ with s then there is an algorithm that solves* identity *problem for class $\mathcal{P}$ with at most $O(s \log s)$ samples.*

If there is an algorithm that solves closeness *problem for class P with at most s samples, then there is an algorithm that solves* distinguishability *problem with at most s samples.*

3 Results

Our algorithmic results can be described as follows:

Theorem 1. *Consider a class of distribution pairs such as $\frac{||P-Q||_2^2}{||P+Q||_2} \geq \alpha$, and let $s = 60(|\log \alpha|)^{7/2}/\alpha$, and each p_i and q_i is at most $\frac{1}{2s}$, then Algorithm 3 produces correct answer with probability at least $1 - c/s^2$, where c is some universal constant.*

Essentially the theorem abouve states that $\frac{||P+Q||_2}{||P-Q||_2^2}$ controls the distinguishability of distribution pair. There are several interesting cases, if $||P - Q||_2$ is comparable to either $||P||_2$ or $||Q||_2$, then the results says that $s \approx 1/||P - Q||_2$ suffices. This generalizes the observation from [2] that if the L_2 difference is large then constant number of samples suffices.

We also note that the condition $p_i \leq \frac{1}{s}$ is a technical condition that guarantees that any fixed element has expectation of appearance of at most $1/2$. In other words, no elements can be expected to be seen in the sample with high probability. This is requirement is particularly striking, given that the results of [10] say that elements that have low expectation, are provably *n*ot useful

when testing continuous properties. Further exploiting this contrast is a very interesting open direction.

For the lower bound part, our results apply to a special class of distributions that we call *w*eakly disjoint distributions

Definition 1. *Distributions P and Q are weakly disjoint if every element x satisfies one of the following:*

$$(1)\ p_x = q_x, (2)\ p_x > 0 \text{ and } q_x = 0, (3)\ q_x > 0 \text{ and } p_x = 0$$

We denote the set of elements such that $p_x = q_x$ by $\mathfrak{C}(P,Q)$, and the rest are denoted by $\mathfrak{D}(P,Q)$

It is worth noting that the known worst case examples of [2] belong to this class. We conjecture that weakly disjoint distributions represent the *hardest* case for lower bound, and all other distributions need fewer samples, and that the result above generalizes to all distributions.

For the rest of the paper we concentrate on the distinguishability problem, but results through lemma 1 immediately apply to closeness problem. Now we formulate our lower bounds results:

Theorem 2. *Let P and Q are weakly disjoint distributions, and let*

$$s \le \min\{\frac{0.25}{||P-Q||_3}, 1/||P||_\infty, 1/||Q||_\infty, \frac{1}{c}\frac{||P+Q||_2}{||P-Q||_2^2}\}$$

where c is some universal constant. No algorithm can solve a distinguishability problem for a class of distritribution pairs defined by arbitrary permutations of (P,Q), e.g. $\mathcal{P} = \{((\pi\mathcal{P}, \pi\mathcal{Q})\}$.

The first, second and third constraints in min expression above are technical assumptions. In fact for many distributions including the worst-case scenario of [2] $\frac{||P+Q||_2}{||P-Q||_2^2} << \min\{1/||P-Q||_3, 1/||P||_\infty, 1/||Q||_\infty\}$, and hence those constraits can be dropped.

4 Preliminaries

4.1 Distinguishing between sequences of Bernoulli random variables

Consider two sequences of random Bernoulli variables x_i and y_i. In this section we characterize when the sign of the observations of $\sum x_i - \sum y_i$ can be expected to be the same as sign of $\mathrm{E}[\sum x_i - \sum y_i]$ with high probability. We defer the proof to the appendix.

Lemma 2. *Suppose $\{x_i\}$ and $\{y_i\}$ is a sequence of Bernoulli random variables such that $\mathrm{E}[\sum x_i] = \alpha$ and $\mathrm{E}[\sum y_i] = \beta$, where $\alpha < \beta$. Then*

$$\Pr\left[\sum x_i > \sum y_i\right] < 2\exp[-\frac{(\alpha-\beta)^2}{8(\alpha+\beta)}]$$

4.2 Weight Concentration Results For Heterogeneous Balls and Bins

Consider a sample $\mathcal{S} = \{s_1, s_2, \ldots, s_s\}$ of size s from a fixed distribution P over the domain $[n]$. Let the sample configuration be $\{c_1, \ldots c_n\}$, where c_i is the number of times element i was selected. A standard interpretation is that sample represents s balls being dropped into n bins, where P describes the probability of a ball going into each bin. The sample configuration is the final distribution of balls in the bins. Note that $\sum c_j = s$. We show tight concentration for the quantity $\sum_{i=1}^{n} \alpha_i c_i$, for non-negative α_i.[1]

Note that c_i and c_j are correlated and therefore the terms in the sum are not independent. An immediate approach would be to consider a contribution of ith sample to the sum. It is easy to see that the contribution is bounded by $\max \alpha_i$, and thus one can apply McDiarmid inequality[8], however the resulting bound would be too weak since we have to apply an uniform upper bound.

In what follows we call the sampling procedure , where each sample is selected independently from distribtuion over the domain $[n]$ the *type I* sampling. Now, we consider a different kind of sampling that we call type II.

Definition 2 (Type II sampling). *For each i in $[n]$ we flip p_i-biased coin s times, and select the corresponding element every time we get head.*

We show that for almost any sample selection of type (I), the corresponding configuration in type II sampling would have similar weight in type II. The weight of all configurations in type (I) not satisfying this constraint is $o(1/s)$ ε_1. Once we show that, then any concentration bound in Type II will translate to corresponding concentration bound for type I.

We use $P^{(I)}[\cdot]$ and $P^{(II)}[\cdot]$ to denote probability according to type (I) and type (II) sampling. Due to space limitations all the proofs are deferred to appendix. Now we show the lower and upper bound connections between type I and type II sampling.

Lemma 3. *For every configuration $\mathcal{C} = \{i_1, \ldots i_n\}$, such that $i_j \leq \ln s$, and $\sum i_j = s'$, where $s' \in [s \pm \sqrt{s}]$*

$$\frac{2}{3}\sqrt{s} P^{(II)}[\mathcal{C}] \leq P^{(I)}[\mathcal{C}] \leq 30 s^{3/2} P^{(II)}[\mathcal{C}].$$

where $P^{(I)}[\mathcal{C}]$ is a probability of observing $\mathcal{C}$ in type I sampling with s' elements and $P^{(II)}[\mathcal{C}]$ is a probability of observing Cin type II sampling with s elements.

The following lemma bounds the total probability of a configuration which elements appearing more than $\ln s$ times. The proof is deferred to appendix.

Lemma 4. *In type I sampling, the probability there exists an element that was sampled more than $\ln s$ times is at most $\frac{1}{s}^{\ln \ln s}$.*

[1] In fact the technique applies to arbitrary bounded functions.

Now we formalize the translation between concentration bounds for Type (I) and Type (II) samplings.

Lemma 5. *Suppose we sample from distribution P, s times using type I and type II sampling resulting in configurations C. Let $A = \{\alpha_i\}$ be an arbitrary vector with non-negative elements, and $r \geq 0$. Then*

$$P^{(I)}\left[|W - \mathrm{E}\left[W\right]| > r\right] \leq 30 s^{3/2} P^{(II)}\left[|W - \mathrm{E}\left[W'\right]| > r\right] + \frac{1}{s^{\ln \ln s}},$$

where $W = \sum_{i=1}^{n} \alpha_i c_i$.

Lemma 6. *Consider s samples selected using Type I sampling from the distribution $P = \{p_i\}$, where $s \geq 10$, and $A = \{\alpha_i\}$ is arbitrary vector. Let $W = \sum_{i=1}^{n} \alpha_i c_i$. Then*

$$\Pr\left[|W - \mathrm{E}\left[W\right]| \geq 2(\ln s)^{3/2} ||A||_2\right] \leq \frac{1}{s^2}$$

5 Algorithm and Analysis

At the high level our algorithm works as follows. First, we check if the two-norms of distributions P and Q are sufficiently far away from each other. If it is the case, then we can decide simply by looking at the estimates of the 2-norm of P, Q and T. On the other hand if $||P||_2 \approx ||Q||_2$ then we show that counting the number of collisions of sample from T, with P and Q, and then choosing the one that has higher number of collisions gives the correct answer with high probability. Algorithm 2 provides a full description on 2-norm estimation. The idea is to estimate a probability mass of a sample S by computing the number of collisions of fresh samples with S, and then noting that the expected mass of a sample of size l is $l||P||_2$. One important caveat is that if S contains a particular element more than once, we need to carefully compute the collisions in such a way to keep the probability of a collision at $l||P||_2$ and to achieve that, we split our sampling into $\max c_i$ phases. During phase i we only count collisions with elements that have occurred at least i times. For more details we refer to algorithm 1 which is used as subroutine for both 2 and the main algorithm 3. For the analysis we mainly use the technique developed in Section 4.

Lemma 7. *Algorithm 1 outputs a set, S, such that* $\mathrm{E}\left[|S|\right] = \sum_{i=1}^{n} c_i p_i$.

Proof. Let w_i be the contribution out of s_i. $\mathrm{E}\left[w_i\right] = \sum_{j=1}^{n} p_j \mathbf{1}_{\mathbf{c_j} \geq \mathbf{i}}$ Summing over all elements of S and use linearity of expectation we have: $\mathrm{E}\left[|S|\right] = \sum_{i=1}^{m} \mathrm{E}\left[w_i\right] = \sum_{i=1}^{m} \sum_{j=1}^{n} p_j \mathbf{1}_{\mathbf{c_j} \geq \mathbf{i}} = \sum_{i=1}^{n} p_i c_i$, where the last equality follows from changing the summation order. ∎

Lemma 8. *The total number of elements selected after l iterations in step 3 is a sum of independent Bernoulli random variables, and its expectation is lW.*

Sampling according to given pattern

Algorithm 1 Sampling according to given pattern

Input: Configuration $\{c_1, \ldots c_n\}$, where $m \geq c_i \geq 0$. Distribution P
Output: Multi-Set of elements S, such that $\mathrm{E}\left[|S|\right] = \sum_{i=1}^{n} c_i p_i$
Description:

1. Sample m elements from P, $s_1, \ldots s_m$
2. For each s_i, if $c_{s_i} \geq i$ include s_i into set S

Algorithm 2 Computing 2-norm of distribution P

Input: Distribution P, accuracy parameter l
Output: Approximate value of $||P||_2^2$
Description:

1. Select l samples from P, let $c_1, \ldots c_n$ be the configuration. Note that the expected weight W of the configuration is $l||P||_2^2$
2. If $\max c_i \geq \log l$ report failure
3. Sample with repetition using Algorithm 1 l times. Let r_i be the respective set size from ith simulation. Note that $\mathrm{E}\left[r_i\right] = W$
4. Report $\sum_{i=1}^{l} r_i/l^2$ as the approximate value

Proof. Indeed, the expected number of selections for every invocation Algorithm 1 is W, and that in itself is a sum of $m \leq \log s$ Bernoulli variables. Thus, we have $lt \log s$ Bernoulli random variables with the total expected weight of lW. ■

Furthermore, the total number of samples used by the algorithm is bounded.

Lemma 9. *The total number of samples used by Algorithm 2 is at most* $2l \log l$.

Now we are ready to prove the main property of the algorithm 2.

Lemma 10 (Concentration results for 2-norm estimation). *Suppose the Algorithm 2 is run for distributions P and T with parameter $l > 10$. If $P = T$ then the estimate for $||P||_2$ is greater than estimate for $||T||_2$ with probability $1/2$. If*

$$s(||T||_2^2 - ||P||_2^2) \geq 4(\ln l)^{3/2}(||P||_2 + ||T||_2)$$

then the estimate for $||P||_2$ is smaller than the estimate for $||T||_2$ with probability at least $1 - c/l^2$, where c is some universal constant.

Proof. The first part is due to the symmetry. For the second part, we use Lemma 6 to note that the weight of selection $W_T \leq s||T||_2^2 - 3(\ln l)^{3/2})||T||_2$ or $W_P \geq s||P||_2^2 + 3(\ln l)^{3/2}||P||_2$ with probability at most $\frac{2}{l^2}$. Therefore with probability at least $1 - \frac{2}{l^2}$ we have $|W_T - W_P| \geq (\ln l)^{3/2}(||P||_2 + ||T||_2)$. Using lemmas 8 and 2 we have:

$$\Pr\left[\tilde{W}_T \leq \tilde{W}_P\right] \leq o(\frac{1}{l^2}) + 2\exp\left[\frac{l^2(W_T - W_P)^2}{8(W_T + W_P)}\right] \leq \frac{1}{2l^2} + 2\exp\left[-\frac{l^2(\ln^3 l)(||P||_2 + ||Q||_2)^2}{8l^2(||T||_2^2 + ||P^2||_2)}\right] \leq \frac{1}{l^2},$$

Bringing it all together: main algorithm and analysis

Algorithm 3 Distinguishing between two distributions

Input: Blackbox providing samples from P, Q and T. Estimate $s = \frac{||P+Q||_2}{||P-Q||_2^2}$.[2] *Output:* "P" if T is P and "Q" otherwise
Description:

1. Compute L_2 norm of P, Q and T using Algorithm 2, using accuracy parameter $l = 30s(\ln s)^{3/2}$. Repeat $\log s$ times. Let $\tilde{P}_i$, $\tilde{Q}_i$ and $\tilde{T}_i$ denote the estimated norms in ith iteration.
2. If $\tilde{T}_i \geq \tilde{P}_i$ for all i or $\tilde{T}_i \leq \tilde{P}_i$ for all i then report "Q" and quit.
3. If $\tilde{T}_i \geq \tilde{Q}_i$ for all i or $\tilde{T}_i \leq \tilde{Q}_i$ for all i then report "P" and quit.
4. Else:
   (a) Training Phase: sample $l = 30(\ln s)^{3/2}s$ elements from distributions P and Q. Let C_P and C_Q denote the configuration of elements that were selected.
   (b) Testing Phase: use Algorithm 1 to sample with repetition s times, on both C_P and C_Q using fresh sample each time. Let c_P and c_Q denote the total size of the Algorithm 1 output for C_P and C_Q respectively.
   (c) If $c_P > c_Q$ report "P", otherwise report "Q".

---

where we have used $\exp(-\ln^3 l) < \frac{1}{2(l^2)}$, for $l > 10$. ■

Let W_P and W_Q denote the total probability mass of sample selected by P and Q in distribution T. In other words $W_P = \sum_{i=1}^{n} t_{s_P(i)}$, where $s_P(i)$ is the i-th sample from P.

Now, consider H_1 (e.g. $T = P$). We have: $\mathrm{E}[W_P] = s\sum_{i=1}^{n} p_i^2$ and $\mathrm{E}[W_Q] = s\sum_{i=1}^{n} p_i q_i$, where i-th term is expected contribution of i-th element of the distribution into the total sum for a single selection. Therefore we have

$$\mathrm{E}[W_P - W_Q] = s\sum_{i=1}^{n} p_i(p_i - q_i)$$

Similarly in hypothesis H_2 we have:

$$\mathrm{E}[W_Q - W_P] = s\sum_{i=1}^{n} q_i(q_i - p_i)$$

The rest of the analysis proceeds as follows, we first show that if the algorithm passes the first stage then with high probability $|(||P||_2^2 - ||Q||_2^2)| \leq 4(\ln s)^{3/2}||P+Q||_2$, in which case the majority voting on collision counts gives the desired result. The following lemma is almost immediate from Lemma 10

Lemma 11 (Correctness of the case when $||P||_2 \not\approx ||Q||_2$). *If*

$$s|\sum_{i=1}^{n}(p_i^2 - q_i^2)| \geq \ln^{3/2}(l)(||P||_2 + ||Q||_2)$$

the algorithm terminates on or before step 3 of algorithm 3 with probability at least $1 - c/s^2$*. Further, the probability of incorrect answer if it terminates is at most* $\frac{c}{s^2}$*, for some constant* c*.*

Proof. Without loss of generality we can assume $T = P$. Using the result of lemma 10, the probability that $\tilde{P}_i < \tilde{T}_i$ or for all i is at most $(\frac{1}{2})^{2\log s} \le \frac{1}{s^2}$. Thus probability of incorrect answer is at most $\frac{1}{s^2}$.

Second, if the condition of the lemma satisfied, then from lemma 10 and the union bound the probability of inconsistent measurements is at most $\frac{\log l}{l^2} \le \frac{1}{s^2}$. ■

Given the lemma above, we have that if the algorithm passed stage 1, then

$$-4(\ln l)^{3/2}(||P||_2 + ||Q||_2) \le l(\sum p_i^2 - \sum q_i^2) \le 4(\ln l)^{3/2}(||P||_2 + ||Q||_2)$$

Therefore in hypothesis H_1:

$$\mathrm{E}\left[W_P - W_Q\right] = l(\sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} q_i p_i) \ge l \times \left[\frac{||P-Q||_2^2}{2} - \frac{4(\ln l)^{3/2}}{l}(||P||_2 + ||Q||_2)\right]$$

We have $\ln l = \ln(s \ln^{3/2} s) \le \ln\ln s + \log s + 3 \le 1.5 \log s$, and thus $l \ge 20(\ln l)^{3/2} \frac{||P+Q||_2}{||P-Q||_2^2}$. Substituting we have:

$$\mathrm{E}\left[W_P - W_Q\right] \ge l\left[\frac{||P-Q||_2^2}{2} - \frac{4}{20}||P-Q||_2^2\right] \ge 6(\ln l)^{3/2}(||P||_2 + ||Q||_2)$$

Applying Lemma 6 with weight function P, the probability that either W_P and W_Q deviates from its expectation by more than $2(\ln l)^{3/2}||P||_2$ is at most $\frac{c}{s^2}$ for a fixed constant c. Thus

$$\Pr\left[W_P - W_Q \le (\ln l)^{3/2}(||P||_2 + ||Q||_2)\right] \le \frac{c_2}{s^2} \tag{1}$$

Therefore W_P and W_Q with high probability are far apart, and could be distinguished using the bounds from Lemma 2. Indeed, the total expected number of hits is sW_P and sW_Q, for both P and Q, thus the probability that the total number of hits is for C_P is smaller than C_Q in H_1 is at most:

$$\Pr\left[c_p < c_q\right] \le \exp[-\frac{s(W_P - W_Q)^2}{W_P + W_Q}] \le \exp\left[-\frac{s\ln^3 l||P+Q||_2^2}{l(||P||_2^2 + ||Q||_2^2)}\right] \le \frac{1}{l^2} \tag{2}$$

Combining (1) and (2) we have that for some universal constant c, with probability at least $1 - \frac{c}{s^2}$, C_P will receive more hits than C_Q and symmetrically in H_2, C_Q will receive more hits than C_P. Thus the algorithm produces correct answer with high probability and the proof of Theorem 1 is immediate.

6 Lower Bounds for Weakly Disjoint Distributions

In this Section we prove Theorem 2. First we observe that for any *fixed* pair of distributions, there is in fact an algorithm that can differentiate between them with far fewer samples than our lower-bound theorem dictates[3]. Thus, the main challenge is to prove that there is no universal algorithm that can differentiate between arbitrary pairs of distributions. Here, we show that even the simpler problem where the pair of distributions is fixed, and a random permutation π is applied to both of them, there is no algorithm that can differentiate between πP, and πQ. Since this problem is simpler than the original problem (we know the distribution shape), the lower bound applies to the original problem.

Problem 3 (Differentiating between two known distribution with unknown permutation). Suppose P and Q are two known distributions on domain D. Solve distinguishability problem on the class of distributions defined by $(\pi P, \pi Q)$, for all permutations π.

In Problem 3, the algorithm needs to solve the problem for every π, thus if such algorithm exists, it would be able to solve distinguishability problem with π chosen at random and from the perspective of the algorithm, elements that were chosen the same number of times, are equivalent. Thus, the only factor the algorithm can differentiate upon are counts of how often each element appeared in different phases. More specifically we will use $|(i, j, k)|$ notation to denote the number of elements, which were chosen i-times while sampling from P, j-times while sampling from Q in the training phase and k-times during the testing phase. We also use notation $|(i, j, *)|$ to denote the total number of elements that were selected i and j times during training phase and arbitrary number of times during the testing phase. Finally and $|(i, j, +)|$ to denote the number of elements that were selected at least once during the testing phase. In what follows we use H_1 and H_2 to denote the two possible hypotheses that the testing distribution is in fact P or Q respectively.

To prove this theorem we simplify the problem further by disclosing some additional information about distributions, this allows us to show that some data in the input possesses no information and could be ignored. Specifically, we rely on the following variant of the problem:

Problem 4 (Differentiating with a hint). Suppose P and Q are two known distributions on D and π is unknown permutation on D. For each element that satisfies one of the following conditions the algorithm is revealed whether it belongs to common or disjoint set. (a) selected at least once while sampling from T and at least twice while sampling from P or Q (b) selected at least twice while sampling from P or Q, and belongs to the $\mathfrak{C}(\pi P, \pi Q)$/ The set of all elements for which their probabilities are known is called *hint*. Given the hint, differentiate between πP and πQ.

[3] For instance distinguishing between two uniform distribution on half domain a constant number of samples is sufficient

Note that Problem 4 is immediately reducible to Problem 3, thus a lower bound for 4 immediately implies a lower bound for 3. If an element from the disjoint part of P and Q has its identity revealed, then the algorithm can immediately output the correct answer. We call such elements *helpful*. We call other revealed elements *unhelpful*, note that the set that the unhelpful elements is fully determined by the training phase (these are the elements that were selected at least twice during training phase and belong to the common set). First, we prove the bound on the probability of observing helpful elements. Later we show that knowledge of unhelp does not reveal much to the algorithm.

Lemma 12. *If the number of samples $s \leq \frac{0.25}{||P-Q||_3}$, then the probability that there is one or more helpful elements is at most $\frac{1}{20}$.*

Proof. For every element that has probability of selection during testing phase p, the probability that it becomes helpful is at most $\binom{2s}{3}p^3 < \frac{8s^3p^3}{6} \leq 1.5s^3p^3$ if it belongs to disjoint section, and 0 otherwise. Therefore, the total expected number of helpful elements is $1.5s^3||P-Q||_3^3$. Using Markov inequality we immediately have that probability of observing one or more helpful element is at most $\frac{1}{1.5s^3||P-Q||_3^3} \leq \frac{1}{20}$ as desired. ∎

Since the probability of observing helpful element is bounded by $\frac{1}{20}$, it suffices to prove that any algorithm that does not observe any *helpful* elements, is correct with probability at most $1 - \Omega(1)$. The next step is to show that disclosed elements from the common part are ignored by the optimal algorithm.

Lemma 13. *The optimal algorithm does not depend on the set of unhelpful elements.*

Proof. More formally, let $\mathcal{C}$ denotes the testing configuration, and let Y denotes the unhelpful part of hint. Let $\mathcal{A}(C, Y)$ is the optimal algorithm that takes as that outputs H_1 or H_2. Suppose there exists Y' and Y'' such that $\mathcal{A}(C, Y') \neq \mathcal{A}(C, Y'')$, and without loss of generality we assume $\mathcal{A}(C, Y') = H_1$ and $\mathcal{A}(C, Y'') = H_2$. Of all optimal algorithms we chose the one that minimizes the number of triples (C, Y, Y') that satisfy this property. Without loss of generality $\Pr[C|H_1] \geq \Pr[C|H_2]$ and let $\mathcal{A}_1$ be the modification of $\mathcal{A}$ such that $\mathcal{A}_1(C, Y'') = H_1$. But then the total probability of error for $\mathcal{A}_1$ will be the $\Pr[C|H_1]\, PrY'' - \Pr[C|H_2]\Pr[Y'']$, which is smaller or equal than $\mathcal{A}$ contradiction with either optimality or minimility of $\mathcal{A}$. ∎

So far we showed that *helpful* elements terminate the algorithm, and *nonhelpful* do not affect the outcome. Therefore, the only signatures (i, j, k) that the algorithm has knowledge of will belong to the following set:

$$\{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(0,1,1),(1,0,1),(0,0,2),(2,0,0),(0,2,0)\}.$$

Consider the following random variables $|(0,1,*)|$, $|(1,0,*)|$, $|(0,0,*)|$, $|(0,2,*)|$,$|(2,0,*)|$. They are fully determined by the training phase and thus are independent of the hypotheses. We call these random variables the training configuration. The following lemma is immediate and the proof is deferred to appendix.

Lemma 14. *If the training configuration is fixed then the values* $|(0,1,1)|$, $|(1,0,1)|$, $|(0,0,1)|$ *fully determine all the data available to the algorithm.*

Therefore for fixed configuration of the training phase, the output of optimal algorithm only depends on $|(0,1,1)|$, $|(1,0,1)|$, $|(0,0,1)|$. Consider $h(i,j)$ the total number of elements that were selected during testing phase, and that have signature $(i,j,*)$. Note that $h(i,j) = \sum_{k=1}^{s} k \times |(i,j,k)|$.

Now we prove that no algorithm by observing $h(0,1)$, $h(1,0)$, $h(0,0)$ can have error probability less than 1/17. Again we defer the proof to the appendix due to space constraints.

Lemma 15. *Let* C_P, D_P *and* C_Q, D_Q *be the probability masses of subsets of* $\mathfrak{C}(\pi P, \pi Q)$ *and* $\mathfrak{D}(\pi P, \pi Q)$ *that were sampled during training phase of* P *and* Q *respectively. Assume without loss of generality* $\Pr[D_P \geq D_Q] \geq 1/2$. *Then with probability at least* $1/2$, *in hypotheses* H_1 *for observed* $h(1,0) = a$, $h(0,1) = b$ $h(0,0) = c$, *we have:*

$$\frac{\Pr[\ h(1,0) = a, h(0,1) = b, h(0,0) = c | H_1]}{\Pr[\ h(1,0) = a, h(0,1) = b, h(0,0) = c | H_2]} \leq 8 \tag{3}$$

From this lemma it immediately follows that any algorithm will be wrong with probability at least 1/16. Now we are ready to prove the main theorem of this section.

Proof of theorem 2. By lemmas 13 and 14, the optimal algorithm can ignore all the elements that has signatures other than $|(0,1,1)|$, $|(1,0,1)|$ and $|(0,0,1)|$. Now, suppose there exists an algorithm $\mathcal{A}(x,y,z)$ that only observes $|(0,1,1)|$, $|(1,0,1)|$, $|(0,0,1)|$, and has error probability less then 1/100. Then the algorithm for $|(0,1,+)|$, $|(1,0,+)|$ and $|(0,0,+)|$ that just substitutes the latter into former, will have error probability at most $1/100 + \frac{1}{20} < 1/17$. Indeed, let $x = |(0,1,+)|$, $y = |(1,0,+)|$ and $z = |(0,0,+)| - 2(s - x - y)$ and execute the algorithm $\mathcal{A}(x,y,z)$ will have error at most $1/100 + \frac{1}{20}$. Thus, we contradicted Lemma 15. ■

7 Conclusions and Open Directions

Perhaps the most interesting immediate extension to our work is to incorporate high-frequency elements in our analysis to eliminate the technical assumptions that we make in our theorems theorem. One possible approach is to combine techniques of Valiant and Micali, but it remains to be seen if the hybrid approach will produce better results. Proving or disproving our conjecture that weakly disjoint distribution are indeed the hardest when it comes to telling distribution apart would also be an important technical result.

The other directions is to extend the techniques of section 4. For instance this techniques could be used to estimate various concentration bounds on how many heterogeneous bins will receive exactly t balls and it remains an interesting question on how far those techniques could be pushed. On the more technical

side an interesting question is whether (under some reasonable assumption), the probability ratio between type I and type II configurations is, in fact bounded by constant, rather than by $O(s)$. By using tighter analysis we can in fact show that this ratio is in fact $O(\sqrt{s})$ though reducing it further remains an open question.

References

1. T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *Proceedings 2001 IEEE International Conference on Cluster Computing*, pages 442–451, 2003.
2. T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that distributions are close. *Foundations of Computer Science, Annual IEEE Symposium on*, page 259, 2000.
3. Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The Complexity of Approximating the Entropy. *Computational Complexity, Annual IEEE Conference on*, 0:17, 2002.
4. Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing - STOC '04*, page 381, 2004.
5. S. Ben-David, J. Gehrke, and D. Kifer. Detecting change in data streams. In *International Conference on Very Large Data Bases*, pages 181–190, 2004.
6. Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.
7. Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58:301:13–30, March 1963.
8. C. McDiarmid. On the method of bounded differences, Surveys in combinatorics, 1989 (Norwich, 1989). *London Math. Soc. Lecture Note Ser*, 141:148–188, 1989.
9. D. Ron. Property testing (tutorial). In *Handbook of Randomization*. 2000.
10. P.P.A. Valiant. Testing symmetric properties of distributions. In *STOC*, pages 383–392, New York, NY, USA, 2008. ACM.

A Equivalence of identity and distinguishability problems

Proof of Lemma 1. We first show how to simulate a correct algorithm $\mathcal{A}_\mathcal{I}$ for the identity problem using an algorithm $\mathcal{A}_\mathcal{D}$ for the distinguishability problem. We run $\mathcal{A}_\mathcal{D}$ for $3 \log s$ times on fresh input, where the sample for the testing phase is always taken from the first distribution. If the answer is always the same say "different", otherwise say the "same". If the two distributions are the same, then $\mathcal{A}_\mathcal{D}$ gives random answer, therefore the probability for $\mathcal{A}_\mathcal{D}$ produce the same answer for $\log s$ iterations (and hence $\mathcal{A}_\mathcal{I}$ producing wrong answer) is at most $(3/5)^{3 \log s} \leq \frac{1}{s^2}$. Similarly if distributions are different, the probability for $\mathcal{A}_\mathcal{D}$ go give different answer means that it has mistaken at least once, which would happen with probability at most $\frac{\log(3s) polylog(s)}{s^2}$ as desired as it remains polylograthmic.

For the other direction we simulate a correct algorithm for the distinguishability problem using an algorithm $\mathcal{A}_\mathcal{I}$ for the identity problem. We test a new

sample against both X and Y, if $\mathcal{A}_{\mathcal{I}}$ says that both are the same or both are different, output X with probability 0.5. If the output of $\mathcal{A}_{\mathcal{I}}$ is "the same as X" output "X", otherwise output "Y"'.

If X and Y are the same the testing algorithm will say "the same" with probability at least $1 - 2polylog(s)/s^2$, thus distinguishing algorithm will say X with probability at least $0.5 - polylog(s)/s^2$, and Y with probability at least $0.5 - polylog(s)/s^2$. If X and Y are different, then the testing algorithm will make a mistake with probability at most $2polylog(s)/s^2$, and thus the identity algorithm will produce different answers, and the distinguishing algorithm will be correct with probability at least $1 - 2polylog(s)/s^2$. ■

B Proofs From Subsection 4.1

Lemma 16. *Let $\{x_i\}$ and $\{y_i\}$ be a sequence of N independent, though not necessarily identical Bernoulli random variables such that* $\mathrm{E}\left[\sum x_i\right] = pN$ *and* $\mathrm{E}\left[\sum y_i\right] = qN$ *and $pN < qN$. If $N \geq \frac{8|\log 2\delta|(p+q)}{(p-q)^2}$, then*

$$\Pr\left[\sum x_i \geq \sum y_i\right] \leq \delta.$$

Proof. Using Chernoff inequality for any α we have

$$\Pr\left[\sum_{i=1}^{n} x_i \geq (1+\alpha)pN\right] < \frac{\exp[\alpha \times pN]}{(1+\alpha)^{(1+\alpha)pN}}. \tag{4}$$

On the other hand each for y_i we have:

$$\Pr\left[\sum_{i=1}^{n} y_i \leq (1-\beta)qN\right] < \exp[-\frac{Nq\beta^2}{2}]. \tag{5}$$

Choose $1 - \beta = \frac{p+q}{2q}$, and $1 + \alpha = \frac{p+q}{2p}$. Since $p < q$, $\beta, \alpha > 0$.

First consider the case where $p \leq q/8$. Thus substituting it into (4) we get:

$$\Pr\left[\sum_{i=1}^{n} x_i \geq \frac{p+q}{2}N\right] \leq \frac{\exp[\frac{q-p}{2}N]}{(\frac{p+q}{2p})^{N\frac{p+q}{2}}} \leq \exp\left[N\frac{q-p}{2} - \frac{3(p+q)}{4}N\right] \leq \exp[-N(\frac{q}{4} + \frac{5p}{4})]$$
$$\leq \exp[-N\frac{(q-p)}{4}] \leq \exp\left[-N\frac{(q-p)^2}{4(p+q)}\right]$$

where we have used $(p+q)/(2p) \geq 4.5 \geq e^{1.5}$. Similarly for (5) we have:

$$\Pr\left[\sum_{i=1}^{n} y_i \leq (1-\beta)qN\right] \leq \exp\left[-N(p-q)^2/8q\right] \leq \left[-N\frac{(q-p)^2}{8(p+q)}\right] \tag{6}$$

If, on the other hand $q \geq p \geq q/8$, then $\alpha = (p+q)/2p - 1 \leq 4$, and we can use the following variant of Chernoff bound:

$$\Pr\left[\sum_{i=1}^{n} x_i \geq N\frac{p+q}{2}\right] \leq \exp[-pN\alpha^2/4] \leq \exp[-N(q-p)^2/(4p)] \leq \exp[-N(q-p)^2/2(p+q)] \tag{7}$$

and similarly:

$$\Pr\left[\sum_{i=1}^{n} y_i \le N\frac{p+q}{2}\right] \le \exp[-pN\beta^2] \le \exp[-N(q-p)^2/(2p)] \le \exp[-N(q-p)^2/(p+q)]$$

Combining we have the desired result. ■

Lemma 2 immediately follows from lemma 16

C Proofs for concentration bounds for balls and bins

Proof of Lemma 3. Note that type II sampling is over s elements whereas type I is over s' elements. We have,

$$P^{(I)}\left[\mathcal{C}\right] = p_1^{i_1} p_2^{i_2} \dots p_n^{i_n} \frac{s'!}{i_1! i_2! \dots i_n!},$$

whereas

$$P^{(II)}\left[\mathcal{C}\right] = \prod \frac{s!}{i_j!(s-i_j)!} p_j^{i_j} (1-p_j)^{s-i_j} = \frac{p_1^{i_1} p_2^{i_2} \dots p_n^{i_n}}{i_1! \dots i_n!} \prod \frac{s!}{(s-i_j)!} (1-p_j)^{(s-i_j)}$$

Recalling that $i_j \le \ln s$, we have $s^{i_j} \ge \frac{s!}{(s-i_j)!} \ge (s-\ln s)^{i_j}$, and $\sum i_j = s'$ therefore we have:

$$s'^{s'} e^{s-s'} \ge s^{s'} \ge \prod \frac{s!}{(s-i_j)} \ge (s-\ln s)^{s'} = s^{s'}(1-\frac{\ln s}{s})^{s'} \ge s'^{s'} e^{s-s'} \frac{1}{2s}. \quad (8)$$

where we have used $s^{s'} = s'^{s'}(1+\frac{s-s'}{s'})^{s'}$ and $e^{s-s'} \ge (1+\frac{s-s'}{s'})^{s'} \ge \frac{e^{s-s'}}{2}$. Substituting we have:

$$\frac{P^{(II)}\left[\mathcal{C}\right]}{P^{(I)}\left[\mathcal{C}\right]} \ge \frac{s'^{s'} e^{s-s'}}{(2es)s'!} \prod \frac{(1-p_j)^s}{(1-p_j)^{i_j}} \ge \frac{e^s}{9es\sqrt{s'}} \exp\left[-\sum_{j=1}^{n} p_j(s+1)\right] \ge \frac{1}{10e^2 s^{3/2}}$$

where in the first transition we used the fact that $(1-x/(s+1))^s \ge e^{-x}$ for $x < 1$ and $s > 1$, and the fact that $p_j(s+1) < 1$ and finally Stirling formula $s'! \le 3\sqrt{s'}s'/e^{s'}$. As desired. To get the lower bound we have:

$$\frac{P^{(II)}\left[\mathcal{C}\right]}{P^{(I)}\left[\mathcal{C}\right]} \le \frac{s'^{s'} e^{s-s'}}{s'!} \prod (1-p_j)^{s-i_j} \le \frac{e^s}{2\sqrt{s'}} \prod (1-1/s)^{-i_j} (1-p_j)^s \quad (9)$$

$$\le \frac{3e^s}{2\sqrt{s}} \exp\left[-\sum_{j=1}^{n} p_j s\right] \le \frac{3}{2\sqrt{s}} \quad (10)$$

where in the second transition we used $n! \ge 2\sqrt{n}(n/e)^n$ and $p_j < 1/l$, in the third we have used $(1-1/s)^{-s} < 3$, and $(1-x/s)^s \le e^{-x}$. ■

Proof of Lemma 4. Indeed, let d_i denote the total number of times that sample s_i was repeated in $\mathcal{S}$. Recall that we have $p_i s < 1/2$ for all i, and thus we can just uniformly upper bound it:

$$\Pr\left[d_i \geq \ln s\right] \leq \max_k \Pr\left[c_k \geq \ln s\right] \leq \max_k \frac{\exp\left[\ln s - p_i s\right]}{\left[\frac{\ln s}{p_k s}\right]^{\ln s}} \tag{11}$$

$$\leq \max_k \frac{s(p_k s)^{\ln s}}{s^{\ln \ln s}} \leq \left(\frac{1}{s}\right)^{\ln \ln s + 1}. \tag{12}$$

using union bound over all d_i gives us the desired result. ■

Proof of Lemma 5. Let us denote the set of all configurations satisfying $|\sum_{i=1}^n \alpha_i c_i - \mathrm{E}\left[\sum_{i=1}^n \alpha_i c_i\right]| > r$, by $\mathcal{D}$. For any configuration $C \in \mathcal{D}$ such that $\max c_i \leq \ln s$, we can apply lemma 3. And thus $P^{(I)}\left[C\right] \leq 30 s^{3/2} P^{(II)}\left[C\right]$. In addition the total probability mass of all configurations C such that $\max c_i > \ln s$ is at most $\frac{1}{s^{\ln \ln s}}$. Therefore $P^{(I)}\left[\mathcal{D}\right] \leq 30 s^{3/2} \Pr\left[|W' - \mathrm{E}\left[W'\right]| > r\right] + \frac{1}{s^{\ln \ln s}}$ as desired. ■

Proof of Lemma 6. Consider type II sampling with s samples and let W' be the total selection weight. Note that $\mathrm{E}\left[W\right] = \mathrm{E}\left[W\right]' = s||P||_2^2$. Further, we have $W' = \sum_{i=1}^s c_i' p_i$, where c_i' is the count of how many times i-th element was chosen. The individual terms in the sum are independent, however they are not bounded, so we cannot use Hoeffding inequality[7]. Instead we consider $V' = \sum_{i=1}^s \min(c_i', \ln s) p_i$. Using Hoeffding inequality we have:

$$P^{(II)}\left[|V' - \mathrm{E}\left[V'\right]| \geq 1.5(\ln s)^{3/2}||A||_2\right] \leq \exp\left[\frac{-4.5 \ln^3 s||A||_2^2}{(\ln s)^2 \sum_{i=1}^n \alpha_i^2}\right] \leq \frac{1}{s^4}$$

furthermore, because of the lemma 4 we have:

$$\mathrm{E}\left[V' - W'\right] \leq \Pr\left[V' \neq = W'\right] ||A||_\infty s \leq \frac{||A||_\infty s}{s^{\ln \ln s}} \leq ||A||_2$$

and hence

$$\Pr\left[|W' - \mathrm{E}\left[W\right]'| \geq 2(\ln s)^{3/2}||A||_2\right] \leq \Pr\left[|W' - \mathrm{E}\left[V\right]'| \geq 1.5(\ln s)^{3/2}||P||_2\right] \leq \frac{2}{s^4}$$

where we have used that $V' \neq W' \leq \frac{1}{s^4}$). Thus, using the concentration lemma 5, we have

$$\Pr\left[|W - \mathrm{E}\left[W\right]| \geq 2(\ln s)^{3/2}||P||_2\right] \leq \Pr\left[|W' - \mathrm{E}\left[W\right]'|2(\ln s)^{3/2}||P||_2\right] \times s^{3/2} + \frac{1}{s^{\ln \ln s}} \leq 1/s^2$$

■

D Proofs for lower bounds

Proof of Lemma 14. Indeed:

$|(0,0,2)| = (s - |(1,0,1)| - |(0,1,1)| - |(0,0,1)|)/2$
$|(1,0,0)| = |(1,0,*)| - |(1,0,1)|, \ |(0,1,0)| = |(0,1,*)| - |(0,1,1)|$
$|(0,0,0)| = |(0,0,*)| - |(0,0,2)| - |(0,0,1)|, \quad |(0,2,0)| = |(0,2,*)|, \quad |(2,0,0)| = |(0,2,*)|$

■

Proof of Lemma 15. Denote $\Pr\left[h(1,0) = a, h(0,1) = b, h(0,0) = c | H_i\right]$ as π_i. We need to bound π_1/π_2. We have $\pi_1 = (C_P + D_P)^a C_Q^b (1 - C_P - C_Q - D_P)^c$ and $\pi_2 = (C_P)^a (C_Q + D_Q)^b (1 - C_P - C_Q - D_Q)^c$ therefore

$$\frac{\pi_1}{\pi_2} = \left(1 + \frac{D_P}{C_P}\right)^a \left(1 + \frac{D_Q}{C_Q}\right)^{-b} \left(1 + \frac{D_P - D_Q}{1 - C_P - C_Q - D_P}\right)^{-c} \tag{13}$$

$$\leq 2\exp\left[\frac{D_P}{C_P}a - \frac{D_Q}{C_Q}b - \frac{D_P - D_Q}{1 - C_P - C_Q - D_P}c\right] \tag{14}$$

$$\tag{15}$$

Now, we replace $a = a_0 + \delta_a$ with $b = b_0 + \delta_b$ and $c = c_0 + \delta_c$, where $a_0 = (C_P + D_P)s$, $b_0 = C_Q s$, and $c_0 = (1 - C_P - C_Q - D_P)s$ and we get:

$$\frac{\pi_1}{\pi_2} \leq 2\exp\left[\frac{D_P^2}{C_P}s\right]\exp\left[\frac{D_P}{C_P}\delta_a\right]\exp\left[\frac{D_Q}{C_Q}\delta_b\right]\exp\left[\frac{|D_P - D_Q|}{1 - C_P - C_Q - D_P}\delta_c\right]$$

Note $\mathrm{E}\left[\delta_a\right] = \mathrm{E}\left[\delta_b\right] = \mathrm{E}\left[\delta_c\right] = 0$, and with probability at least 0.5, $\delta_a \leq 3\sqrt{(C_P + D_P)s}$, $\delta_b \leq 3\sqrt{C_Q s}$ and $\delta_c \leq 3\sqrt{(1 - C_P - C_Q - D_P)s}$. Substituting we have with high probability:

$$\frac{\pi_1}{\pi_2} \leq 2\exp\left[\frac{D_P^2}{C_P}s\right]\exp\left[\frac{3D_P\sqrt{s}}{\sqrt{C_P}}\right]\exp\left[\frac{3D_Q\sqrt{s}}{\sqrt{C_Q}}\right]\exp\left[\frac{3|D_P - D_Q|\sqrt{s}}{\sqrt{(1 - C_P - C_Q - D_P)}}\right] \tag{16}$$

Let $||P_D||_2$ denotes the 2 norm of the disjoint part of P and Q , and $||P_C||_2$ the the 2-norm of common part P (or Q). We have $||P - Q||_2^2 = ||P_D||_2^2 + ||P_Q||_2^2$ and $||P + Q||_2 \leq ||P_C + P_D||_2$.

With high probability we have $D_P \leq 2||P_D||_2^2 s \leq \frac{||P_C||_2}{4}$, where we used concentration bounds and substituted $s \leq \frac{1}{10}\frac{||P+Q||_2}{||P-Q||_2^2} \leq \frac{||P_C||_2}{8(||P_D||_2)^2}$. Similarly $D_Q \leq ||P_C||_2/4$ and $C_P \geq \frac{||P_C||_2^2 s}{2}$, therefore we have

$$\frac{D_P^2 s}{C_P} \leq \frac{2||P_C||_2^2 s}{16||P_C||_2^2 s} \leq 1/10$$

Substituting in (16) and using the fact the last exponent is $o(1)$, we get the desired result. ■